Clustering the Health Indicators: A Framework for Evaluating Population Health

Yujia Sun, MS[1], Md Tareq Ferdous Khan, PhD[1], Ronald W. Gimbel, PhD[1], Jihad S. Obeid, M.D[2], Xia Jing, MD, PhD[1]
[1]Department of Public Health Sciences, College of Behavioral, Social and Health Sciences, Clemson University, Clemson, SC; [2] Medical University of South Carolina, Charleston, SC

Background

We identified 29 key health indicators [1,2] from previous research that can be used to construct a Health Indicator Index (HII) for population health assessment. However, many secondary datasets and electronic health records (EHR) lack complete information on all indicators. A broader clustering of these indicators may help overcome this limitation, and thus, this study aims to cluster HII using clustering approaches.

Objectives

The goal of this study was to apply clustering approaches to group the 29 health indicators,[1,2] in order to identify coherent subsets that could be used when complete indicator data is unavailable.

Research Design

Using data from a previous survey on public perception of health indicators, we used hierarchical and K-means clustering approaches to cluster the 29 health indicators. These clustering can be evaluated by cross-referencing the results through overlapping to determine whether clusterings logically align with their real-world relationships.

Results

Based on the dendrogram cutoff of hierarchical clustering,[3] four clusters were identified that best aligned with our understanding of these health indicators: preventative health measures, metabolic health, behavioral and lifestyle factors, social determinants and wellbeing. (Figure 1)

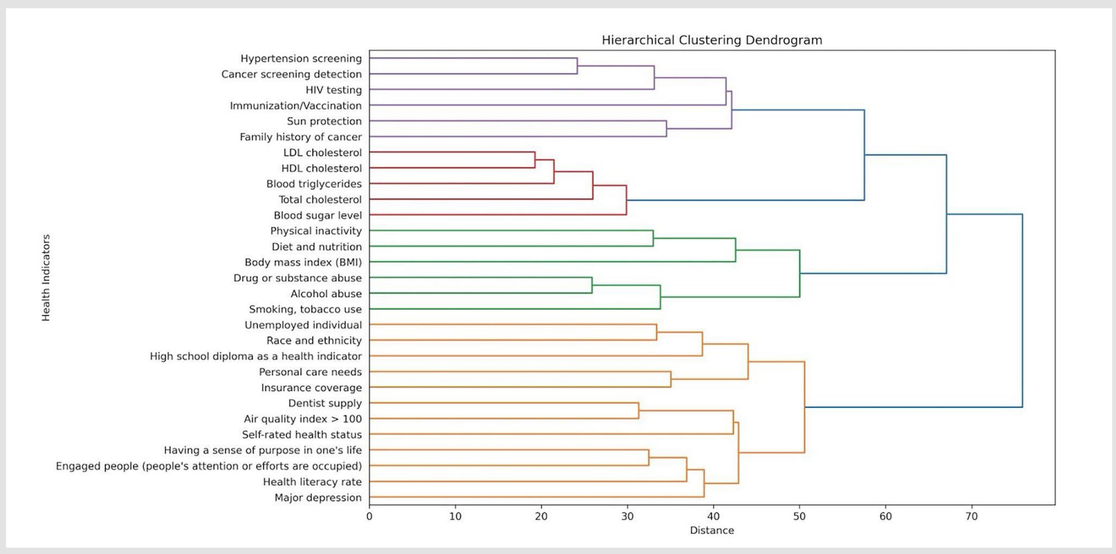

Figure 1. Hierarchical Clustering of 29 Health Indicators

K-means clustering [4] (K=7 based on the elbow method and silhouette score [5]) produced a finer resolution with seven clusters (Figure 2). We found that the cluster separation was not particularly strong, as indicated by the silhouette score equals 0.115. However, the results reinforce commonly seen patterns across methods (Figure 3).

- Metabolic health such as blood sugar, blood triglycerides, HDL cholesterol, LDL cholesterol, and total cholesterol are consistently clustered together, which emphasizes their key role in evaluating cardiovascular and metabolic risks across both methods.
- Preventative health measures such as family history of cancer, sun protection, vaccination, and cancer screening also continue to be grouped together in both clustering methods. This reinforces their shared purpose of early detection and prevention.

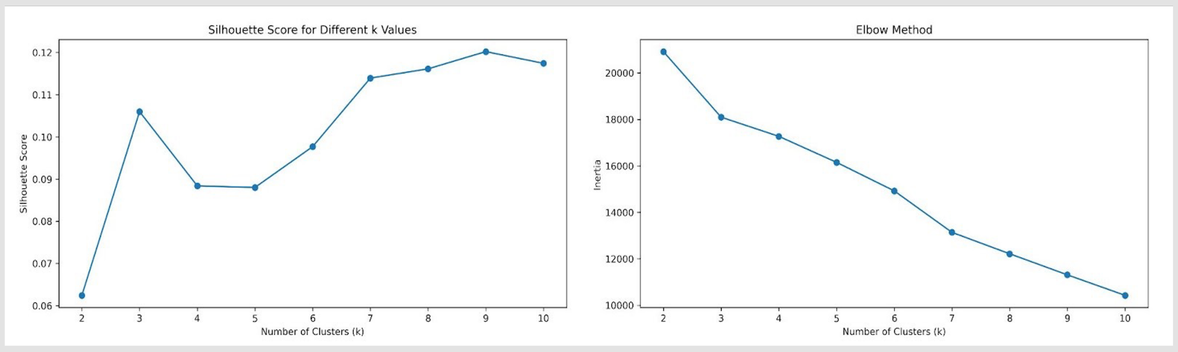

Figure 2. Silhouette (left) and Elbow Method Results (right) for Optimal K in K-means Clustering

Conclusions & Future Plans

This clustering approach may help identify coherent subsets of health indicators, which could later inform the selection of representative variables for a health index when full data is unavailable.

Future research will explore other clustering methods, such as Latent Profile Analysis (LPA), HII composition based on clustering results and HII validation using real-world data. (Figure 4)

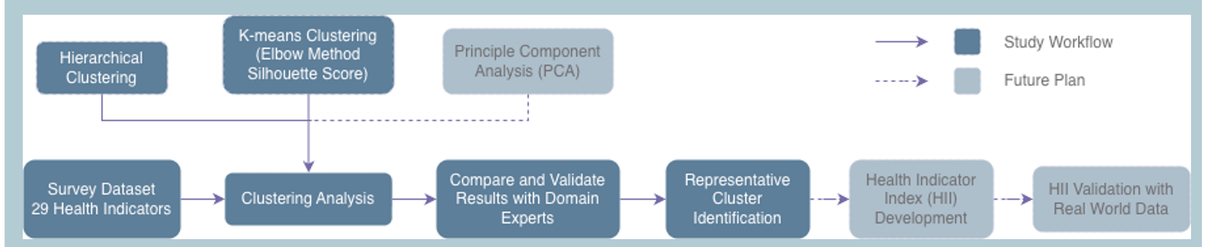

Figure 4. Study Workflow and Future Plan for HII Development

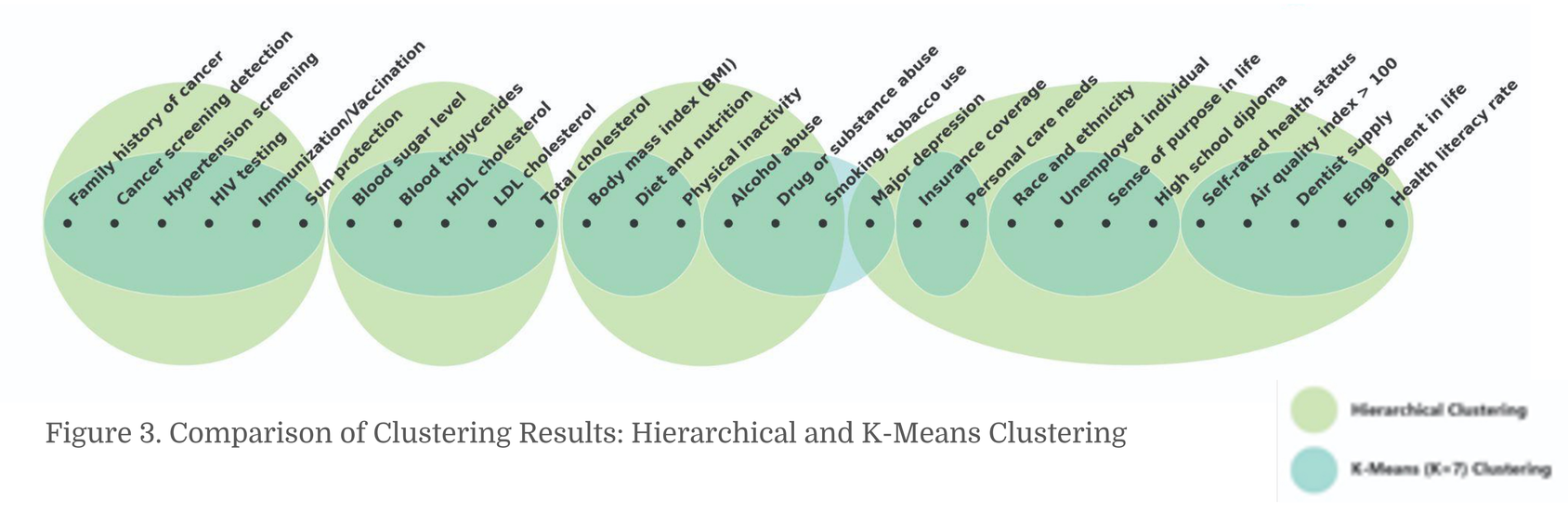

Figure 3. Comparison of Clustering Results: Hierarchical and K-Means Clustering

References

1. Sokoya T, Zhou Y, Diaz S, et al. Health indicators as measures of individual health status and their public perspectives: Cross-sectional survey study. J Med Internet Res. 2022;24(6):e38099. doi:10.2196/38099.
2. Sokoya T, Zhou Y, Diaz S, et al. Health indicators as measures of individual health status, their perceived importance and associated factors—An observational study. JMIR Public Health Surveill. 2025;11:e65616. doi:10.2196/65616.
3. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–244.
4. MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol 1. University of California Press; 1967:281–297.
5. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math. 1987;20(1):53–65.